# Joint Modelling of Longitudinal and Survival Data: A Comparison of Joint and Independent Models

McCrink, Lisa[1]; Marshall, Adele H.[2]; Cairns, Karen[3]
*Queen's University, Belfast, Centre for Statistical Science & Operational Research (CenSSOR)*
*David Bates Building, University Road*
*Belfast BT7 1NN, United Kingdom*
*Email: lmccrink01@qub.ac.uk [1]; a.h.marshall@qub.ac.uk [2]; k.cairns@qub.ac.uk [3]*

## 1. Introduction

In recent years, the interest in longitudinal data analysis has grown rapidly through the development of new methods and the increase in computational power to aid and further develop this field of research. One such method is the joint modelling of longitudinal and survival data.

It is commonly found in the collection of medical longitudinal data that both repeated measures and time-to-event data are collected. These processes are typically correlated, where both types of data are associated through unobserved random effects. Due to this association, joint models were developed to enable a more accurate method to model both processes simultaneously.

When these processes are correlated, the use of independent models can cause biased estimates [3, 6, 7], with joint models resulting in a reduction in the standard error of estimates. Thus, with more accurate parameter estimates, valid inferences concerning the effect of covariates on the longitudinal and survival processes can be obtained. The purpose of this research is to compare independent and joint models to demonstrate the benefits of using joint models to handle correlated longitudinal and survival data.

This research will investigate the known association between haemoglobin fluctuations and the survival of dialysis patients [8]. The focus of this research will be on the survival of haemodialysis patients where the effect of haemoglobin is accounted for as a time-dependent covariate. To do so, joint and independent models will be built using UK Renal Registry (UKRR) data, collected over a three year period, consisting of 5860 haemodialysis patients with just over 59,000 observations.

One approach commonly used to build a joint model is to simultaneously model the longitudinal and survival processes by linking them using unobserved random effects through the use of a shared parameter model. The joint model links a multilevel mixed model and a Cox Proportional Hazards (PH) model. It will be compared to the corresponding independent models.

The remainder of this paper is organised as follows. Section 2 details the joint model structure and Section 3 describes the UK renal data that will be modelled using the shared parameter model. The results of this model and the corresponding independent models will be given and discussed in Sections 4 and 5 respectively.

## 2. The Joint Model Structure

This research focuses on the use of a joint model, where the longitudinal and survival processes are assumed to be conditionally independent given unobserved random effects. This type of joint model is also called a shared parameter model, as both processes share these random effects [9, 10, 11]. The joint model considered in this research links a multilevel mixed model and a Cox PH model, similar to the model proposed by Wulfsohn and Tsiatis [11].

The multilevel mixed model is of the form:

$$
\begin{aligned}
(1) \qquad y_i(t) &= m_i(t) + \varepsilon_i(t) \\
(2) \qquad &= x_i'(t)\beta + z_i'(t)b_i + \varepsilon_i(t) \qquad \text{where } \varepsilon_i(t) \sim \mathrm{N}(0, \sigma^2)
\end{aligned}
$$

where $m_i$ represents the true longitudinal response and $x_i$ and $z_i$ are the design matrices for the fixed and random covariates respectively, with corresponding parameter estimates $\beta$ and $b_i$. The random measurement error term $\varepsilon_i(t)$ is assumed to be independent of $b_i$, where $b_i$ is assumed to follow a multivariate normal distribution with variance $D$ such that $b_i \sim \mathrm{N}(0, D)$.

This multilevel mixed model will be linked to a Cox PH model with an unspecified baseline hazard function, $h_0(t)$. Let $T_i$ represent the failure time for the $i^{th}$ individual such that either censoring or the event has occurred, where $T_i^*$ represents the true event time. The hazard for the $i^{th}$ individual is given by:

$$
\begin{aligned}
h_i(t \,|\, \mathcal{M}_i, w_i) &= \lim_{dt \to 0} \left\{ \frac{\Pr\left[t \le T_i^* < (t + dt) \,|\, T_i^* \ge t, \mathcal{M}_i(t), w_i\right]}{dt} \right\} \\
(3) \qquad &= h_0(t) \exp\{\gamma' w_i + \alpha m_i(t)\}
\end{aligned}
$$

where $\mathcal{M}_i$ represents the history of the true longitudinal response, $m_i(t)$, up to time $t$, $w_i$ represents the baseline covariates, with corresponding parameter estimates $\gamma$, and $\alpha$ is the parameter for the true longitudinal response.

As longitudinal data is typically collected at discrete time points and with error, $m_i(t)$ needs to be estimated to determine the entire history of the longitudinal response, $\mathcal{M}_i(t)$. This is achieved through the multilevel mixed model, Equation 2, and then incorporated into the survival model, Equation 3. Independent models will be built and the parameters obtained will be used as initial values to fit the joint model.

Parameter estimates will be obtained through the use of maximum likelihood estimation. This involves maximising the log-likelihood, given in Equation 4, corresponding to the joint distribution of the longitudinal and survival processes. Both processes share the same unobserved random effects, $b_i$, and are conditionally independent given these random effects.

$$
(4) \qquad \log p(T_i, \delta_i, y_i; \theta) = \log \int p(T_i, \delta_i \,|\, b_i; \theta_t, \beta) \left[ \prod_j p\{y_i(t_{ij}) \,|\, b_i; \theta_y\} \right] p(b_i; \theta_b)\, db_i
$$

where $\theta_t$, $\theta_y$ and $\theta_b$ represents the parameters for the survival process, the longitudinal process and the random-effects covariance matrix respectively, $p\{y_i(t_{ij}) \,|\, b_i; \theta_y\}$ is the density for the longitudinal process and $p(b_i; \theta_b)$ is the density for the random effects. $p(T_i, \delta_i \,|\, b_i; \theta_t, \beta)$ is the likelihood for the survival process given by,

$$
(5) \qquad p(T_i, \delta_i \,|\, b_i; \theta_t, \beta) = [h_i(T_i \,|\, \mathcal{M}_i(T_i); \theta_t, \beta)]^{\delta_i} \; S_i(T_i \,|\, \mathcal{M}_i(T_i); \theta_t, \beta)
$$

where the hazard $h_i(\cdot)$ is given by Equation 3, and, the survivor function for the $i^{th}$ individual is given by,

$$
\begin{aligned}
S_i(T_i \,|\, \mathcal{M}_i(T_i); \theta_t, \beta) &= \Pr(T_i^* > t \,|\, \mathcal{M}_i(T_i), w_i; \theta_t, \beta) \\
&= \exp\left\{ - \int_0^t h_i(s \,|\, \mathcal{M}_i(s); \theta_t, \beta)\, ds \right\} \qquad (6)
\end{aligned}
$$

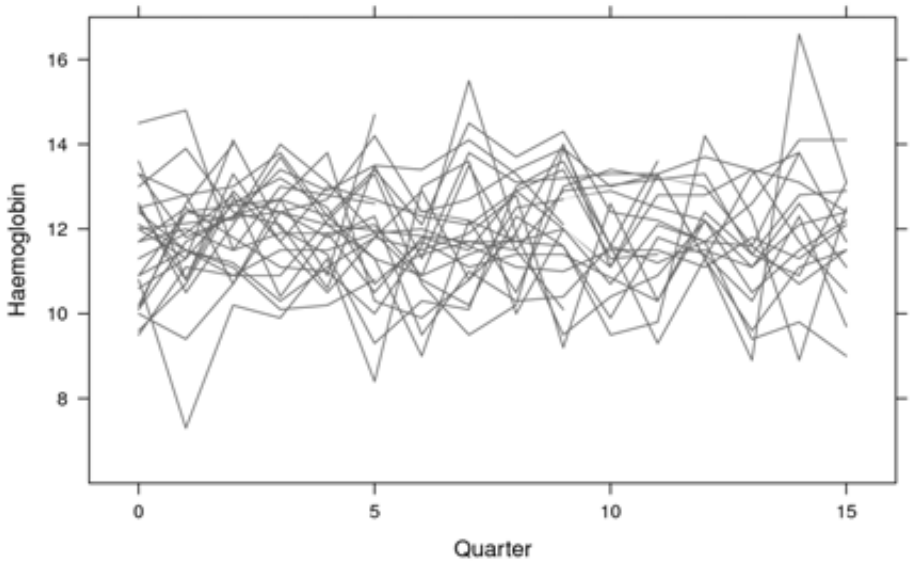
The log-likelihood for the joint model is approximated using the Expectation-Maximisation (EM) algorithm, where the integral in Equation 4 is approximated using the Gauss-Hermite integration rule, as it typically does not have an analytical solution.

## 3. UK Renal Registry Data

The joint model will be built using UKRR data, collected on a quarterly basis over a three year period, commencing at the start of 2005. It consists of clinical and biochemical data for all patients receiving haemodialysis in each renal centre throughout the UK [1]. The sample of data used for this analysis consists of 5860 haemodialysis patients, with a total of 59,168 observations, taken from 67 different renal centres across the UK. This research focuses on the effect of various covariates on haemoglobin variability and patients' survival, including patients' ferritin (iron) levels, cause of death (COD), gender and age.

Haemoglobin variability is clearly demonstrated in Figure 1 and has previously been found to be associated with reduced survival rates for dialysis patients [8]. To further this research, this analysis focuses on the effect of haemoglobin variability on the survival of haemodialysis patients through the use of joint models. Similar methods have previously been used to analyse US renal data [4, 7]. However, due to the differences in end stage renal disease practice patterns, further research is required using UK renal data.

**Figure 1 :** ***Graph illustrating the fluctuations in Haemoglobin over time***



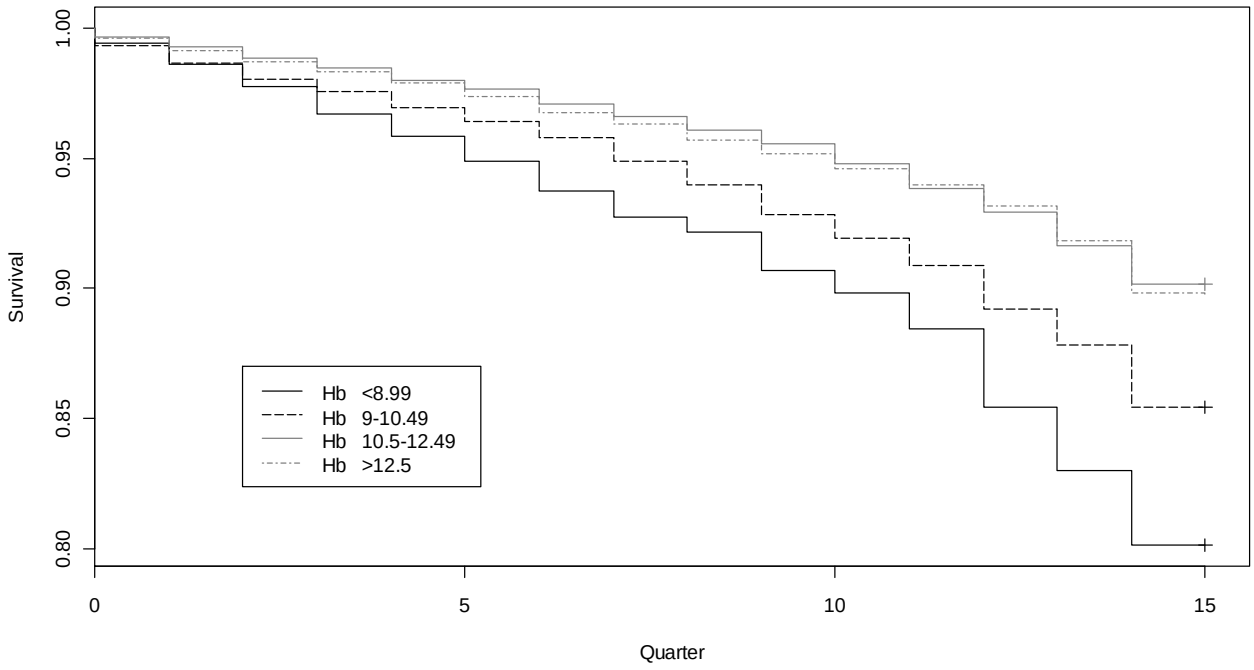

To determine the extent that haemoglobin variability is due to within-individual correlations, the Intraclass Correlation Coefficient (ICC) is calculated. Doing so for the UKRR dataset, the ICC was found to equal 0.25. This indicates that 25% of the variation between individuals can be accounted for due to the dependency of the repeated measurements within individuals. This is taken into consideration through the incorporation of random effects in the longitudinal process.

## 4. Results

A joint model and the corresponding independent models were built using UKRR data to show the benefits of joint modelling when both the longitudinal and survival processes are associated through unknown covariates. The independent multilevel mixed model and a Cox time-dependent model were built using the nlme package [5] and survival package [12] in R respectively.

A time-dependent Cox model was initially built to provide naive estimates of how an individual's haemoglobin levels affect their survival and will be compared to the joint analysis results. The results obtained from the time-dependent Cox model, given in Table 1, agree with previous research [8] and with the preliminary analysis, given in Figure 2, which indicates that patients with higher haemoglobin levels have greater survival rates. This model confirms that haemoglobin is significantly associated with the survival process of haemodialysis patients and thus a shared parameter model is appropriate to analyse this data.

**Figure 2 :** ***Graph illustrating the Kaplan-Meier estimates of survival of haemodialysis patients depending on Haemgoblin level***




The joint model was built using the JM package in R [9] incorporating patients' ferritin (iron) levels into the longitudinal submodel and patients' cause of death (COD) and whether the patient received a transplant prior to the start of the study was incorporated into the survival submodel. The results obtained are also provided in Table 1. The joint model agrees with the preliminary analysis that those patient's with higher haemoglobin levels have a greater survival rate. The significance of the shared parameter that links the two processes, and the reduction in the standard error of the parameter estimates when compared to independent model estimates, indicates the need for a joint analysis of this data compared to the use of independent models.

**Table 1 :** ***Results for the Independent and Joint Models***

| | | **Joint Model** | | **Independent Models** | |
|---|---|---|---|---|---|
| | | Coefficient | Standard Error | Coefficient | Standard Error |
| **Longitudinal Process:** | Intercept | 11.8264 | 0.0159 | 11.8282 | 0.0159 |
| | Time | -0.0161 | 0.0016 | -0.0159 | 0.0016 |
| | Ferritin | -0.0004 | 0.0000 | -0.0004 | 0.0000 |
| **Survival Process:** | COD (CVA) | 2.2192 | 0.1276 | 2.2884 | 0.1291 |
| | COD (Heart) | 2.1754 | 0.0596 | 2.2474 | 0.0623 |
| | COD (Infec) | 2.1300 | 0.0703 | 2.2078 | 0.0623 |
| | COD (Malig) | 2.1985 | 0.1084 | 2.2621 | 0.1101 |
| | COD (Other) | 2.2455 | 0.0885 | 2.3284 | 0.0903 |
| | COD (Trts) | 2.0994 | 0.0718 | 2.1619 | 0.0742 |
| | COD (Uncer) | 1.9681 | 0.0615 | 2.0352 | 0.0638 |
| | Transplant | 0.0883 | 0.0604 | 0.0949 | 0.0605 |
| | Shared Parameter | -0.0156 | 0.0027 | -0.0685 | 0.0126 |

## 5. Conclusion

When the longitudinal and survival processes are correlated, valid inferences can be made through the use of a joint modelling approach. This has been demonstrated using UK renal data to simultaneously model haemodialysis patients' haemoglobin flucatuations over time and their survival. The use of a joint analysis compared to independent models shows a decrease in the standard errors. This reduction in bias means that more accurate inferences can be made using joint model estimates.

Further work is required to determine which parameters are most suitable to model both the longitudinal and survival processes with the aim of aiding clinicians by exploring the causes of haemoglobin variability and its repercussions on haemodialysis patient's survival rates.

However, a major limitation of joint models is the slow convergence rates of these types of models and the large computational power required to fit these models. Future research is therefore needed to improve the estimation techniques for these models.

## 6. Acknowledgements

The authors have collaborated with the UK Renal Registry (UKRR) for this work, who has approved the use of their data for this analysis. In particular, the authors wish to acknowledge Dr Damian Fogarty, director of the UKRR, from Belfast City Hospital. Lisa McCrink is supported by a Department of Employment & Learning (DEL) studentship.

# References

[1] Ansell, D., Castledine, C., Feehally, J., Fogarty, D., Ford, D., Inward, C., Tomson, C., Warwick, G., Webb, L. & Williams, A. 2009, *UK Renal Registry The Twelfth Annual Report.*

[2] Ibrahim, J.G., Chu, H. & Chen, L.M. 2010, "Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data", *Journal of Clinical Oncology*, vol. 28, no. 16, pp. 2796-2801.

[3] Little, R.J.A. & Rubin, D.B. 2002, *Statistical Analysis with Missing Data*, 2nd Edition, Wiley Series in Probability and Statistics, New York.

[4] Liu, L., Ma, J.Z. & O'Quigley, J. 2008, "Joint analysis of multi-level repeated measures data and survival: an application to the end stage renal disease (ESRD) data", *Statistics in medicine*, vol. 27, no. 27, pp. 5679-5691.

[5] Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & the R Core team 2009, "nlme: Linear and Nonlinear Mixed Effects Models".

[6] Prentice, R.L. 1982, "Covariate Measurement Errors and Parameter-Estimation in a Failure Time Regression-Model", *Biometrika*, vol. 69, no. 2, pp. 331-342.

[7] Ratcliffe, S.J., Guo, W. & Ten Have, T.R. 2004, "Joint modeling of longitudinal and survival data via a common frailty", *Biometrics*, vol. 60, no. 4, pp. 892-899.

[8] Gilbertson, D.T., Ebben, J.P., Foley, R.N., Weinhandl, E.D., Bradbury, B.D. & Collins, A.J. 2008, "Hemoglobin level variability: Associations with mortality", *Clinical Journal of the American Society of Nephrology*, vol. 3, no. 1, pp. 133-138.

[9] Rizopoulos, D. 2010, "JM: an R package for the joint modelling of longitudinal and time-to-event data", *Journal of Statistical Software*, vol. 35, no. 9, pp. 1-33.

[10] Sousa, I. 2011, "A Review on Joint Modelling of Longitudinal Measurements and Time-to-event", *REVSTAT*, vol. 9, no. 1, pp. 57-81.

[11] Wulfsohn, M.S. & Tsiatis, A.A. 1997, "A joint model for survival and longitudinal data measured with error", *Biometrics*, vol. 53, no. 1, pp. 330-339.

[12] Therneau, T. & Lumley, T. 2011, "survival: Survival analysis, including penalised likelihood".

## RÉSUMÉ (ABSTRACT) — optional

*Lisa McCrink is a PhD student at Queen's University Belfast, Northern Ireland. Her research focuses on the methodology of models used to handle unstructured longitudinal data, concentrating on medical applications. Lisa holds an MSci in Mathematics and Statistics & Operational Research from Queen's University Belfast and is a member of the Centre for Statistical Science and Operational Research (CenSSOR), Queen's University Belfast.*

*Dr Adele Marshall is a lecturer in Statistics and Operational Research at Queens University Belfast, where she is a Reader in Statistics and the Director of Research at the Centre for Statistical Science and Operational Research (CenSSOR). Adele is President of the Irish Statistical Association (ISA) and her current research focuses on Survival Analysis, Coxian Phase-Type Distributions and data mining.*

*Dr Karen Cairns is also a lecturer in Statistics and Operational Research at Queens University Belfast, where she completed a PhD in Theoretical Physics and her current research focuses on Survival Analysis, Markov Modelling and Longitudinal Data Analysis.*